DOCUMENT RESUME

ED 282 419 FL 016 710

AUTHOR van Roosmalen, Willem M. M.
TITLE Objective Based Tests for Reading Comprehension.
PUB DATE May 83
NOTE 6p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706. Small print in figures may affect legibility.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Grouping; Difficulty Level; Foreign Countries; Item Banks; *Language Tests; Multiple Choice Tests; Objective Tests; *Reading Comprehension; Reading Tests; Secondary Education; Test Construction; *Test Format; Testing Problems; *Test Validity
IDENTIFIERS *Netherlands; *Rasch Model

ABSTRACT

The construction of objective tests for native language reading comprehension is described. The tests were designed for the early secondary school years in several kinds of schools, vocational and non-vocational. The description focuses on the use of the Rasch model in test development, to develop a large pool of homogenous items and establish item difficulty. Each item consists of a passage or text with a maximum of 250 words and one multiple-choice question with four alternative answers. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282419

In this paper I will discuss the construction of some objective based tests for reading comprehension (mother tongue) applying the Rasch model. These tests aim to measure the ability level of students from the first three forms of the several types of secondary schooling, vocational as well as non-vocational, existing in The Netherlands. The students are between 12 and 15 years old. I will focus mainly on the construction proper, but before doing so I shall consider briefly some aspects of the theory of reading comprehension and the reasons for using the Rasch model.

It is only feasible to construct objective based tests for reading comprehension if one assumes that reading comprehension is a set of different skills. The theory is not all too clear about the truth of that assumption. In fact, there is an ongoing debate between those who support this notion and those who assert that reading comprehension is a single global skill not very different from the skill of reasoning. There has been some empirical research intended to try and resolve this problem, but it has not been conclusive. I mention Davis (1944, 1968), Spearrit (1972), Chapman (1971), and Andrich and Godfrey (1979) as advocates of the different skills theory, and Thurstone (1946) and Thorndike (1973) as supporters of the global skill theory. Whether one or the other theory is the right one,is not altogether an academic question. There are some very important implications for the teaching of reading comprehension. If the different skills theory is true, teachers will have to plan carefully the teaching of each skill; if the global skill theory is true,there is no justification for teaching different skills.
As long as theory offers no solution, it seems safest for a teacher to adopt the different skills theory. At least no harm can come from that. And if he does so, objective based tests can be very helpful.
Now the tests we have in mind will not only be objective based, they will also be level tests. It is assumed that the different reading comprehension skills are such that it is unlikely that pupils of 12 - 15 years of age have mastered them completely. They reach a certain ability level and improve on it, but presumably never reach full proficiency.
Bearing this in mind it would be nice if we could construct an instrument with which it would be possible to measure a person's level of performance on a well-defined reading comprehension skill as well as his improvement on that skill.
What we want to measure, however, is not a visible characteristic such as height or weight. It is a mental ability defying direct observation. Such an ability is a latent one. It can, however, reflect itself in observable behaviour. In fact this is just what happens when a

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER(ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL016710

person takes a test. His score on this test is an observable reflection of some latent ability such that we can interprete it as a manifest indicator of that latent ability. To elicit the level of this latent trait from the indicator we need a model that specifies the relation between both. Such a model for instance is the well-known classical test theory. For our purpose however this is not as suitable as we would wish it to be. There are several reasons for this, of which I mention the two more important ones.

(1) In the classical model the testees' scores are test dependent. This entails that we don't know the ability level of the testee as such but only the relation of his ability level to the ability levels of the other testees that took the same test.

(2) The item parameters in the classical test theory are not sample free. So we don't know the difficulty of a specific item as such, but only the difficulty for the sample of testees that took it.

With the Rasch model we don't have these problems: the ability levels of testees are test free and the item difficulties are sample free. It should be clear from this that the Rasch model has some real advantages over the classical model.

One of the more important assumptions in the Rasch model is that all item responses reflect a single latent trait. Now if this should not be true the model is not applicable, so it is absolutely necessary to check on it. To see whether the model fits our particular application, we use the graphical method described in Rasch (1960). I will return to this later on.

When constructing objective based tests, naturally the first thing to do is to collect a set of relevant objectives. And as it is up to the teachers in the field which objectives should be tested, the obvious thing to do is to ask them. This we did and we ended up with a prelimenary set of five objectives:

(1) the student has to identify the main topic of a passage/text;

(2) the student has to identify the main argument of a passage/text;

(3) the student has to identify conclusions;

(4) the student has to identify causes;

(5) the student has to identify effects.

The construction of tests for these objectives was started in 1980. There are none available yet; the first four will be issued in the spring of 1984.

As the ability levels of the students for which the tests are designed vary widely and as we strive at a relatively high precision in the measurement of each student's ability, we need a rather large pool of homogeneous items. In 1980, when we started, we estimated the number of items needed for one test to be about one hundred and at this moment we estimate it to be about ninety.

Each item consists of a passage or text, the length of which is at most some 250 words, and precisely one multiple choice question with four alternatives. These items are constructed by a group of proficient teachers in the field, who have a sound knowledge of at least one of the relevant types of schooling in which the tests will be used. They construct 125 items for each test, as we take into account that some items may turn out to be

bad or misfitting the model. During the construction period all items are thoroughly screened. With the 125 items that in the end are considered to be good items, an initial pretest is run. This is a very complex matter, since we have so many items and such a large range of abilities. Because of the number of items we can not ask each student to answer them all, since it would take too much time. And because of the large range of abilities, we must make sure that the sample - as we can not take the whole population - is such that the estimation of item difficulties is possible on all levels. The sample we use consists of nine groups: the first three forms of each of three types of secondary schooling. We try to get at least 150 students in each group. Each of them is asked to answer 25 items. Ten of these are common items (anchor items), which are used to equate all item difficulties, that is, to put them all on the same scale. The item difficulties are estimated with the help of a Rasch analysis.

The first pretest aims to discover which items are good ones and which are not. We check two things. The student's responses are used to see whether there is anything wrong with the alternatives. Of course we do not need Rasch analysis for this. It is nothing new. However, in addition to this we check if the items fit the model. As I have mentioned before, for this we use the graphical model test described by Rasch (1960).

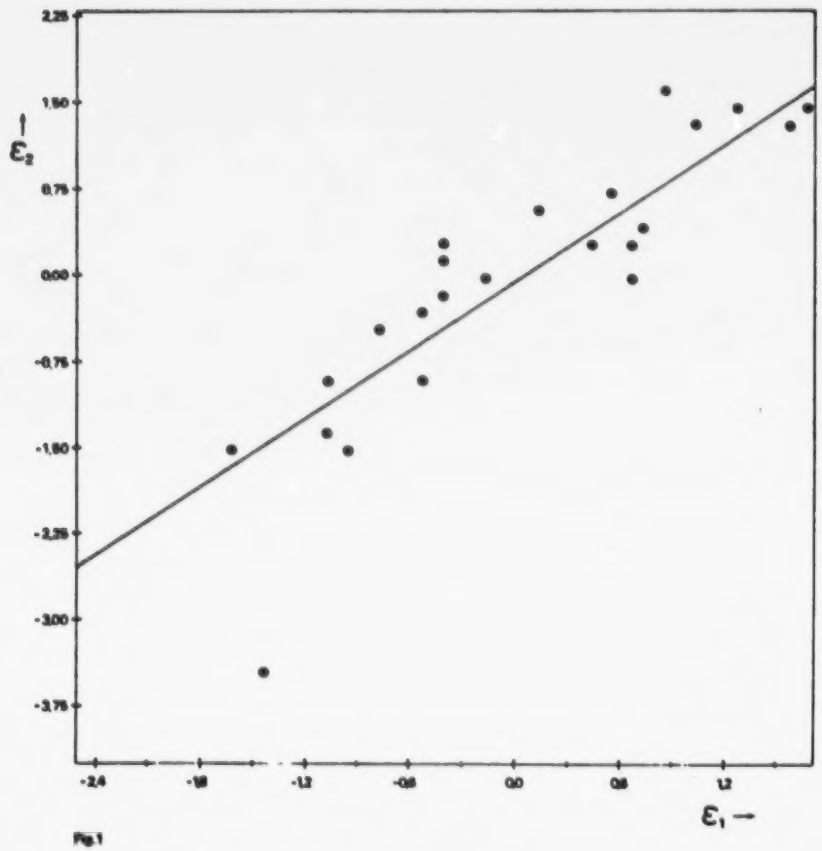

Fig.1

Graphical model fit

Figure 1 gives an example. It shows the testing for model fit of 25 items taken by one of the nine sample groups. The test is based on the assumption of the Rasch model that the item difficulties are sample free. To see whether this is true the sample group is split in two halves. For each half group the item difficulties are calculated. The item difficulties of one half (E_1) are set out on the abscis and of the other (E_2) on the ordinate. Then the items are plotted on the graph according to their two difficulties. In the ideal situation these difficulties should be the same and so the difficulties of the 25 items in the plot should lay on a straight line at an angle of 45°. In reality they appear scattered along this line. As long as the items are relatively close to the 45° line, there is a fair chance that they are unidimensional, i.e. that they measure the same latent trait. It should be stressed, however, that there is no exact criterion for when the item is so far away from this line that it can be said that it does not measure the same trait as the other items. Only one thing is certain, and that is that the further away it is, the less likely it is that it measures the same trait. In using the model test we have to decide what distance from the ideal line we consider to form a convenient boundary. In the case represented by figure 1 we discarded the item in the lower left hand corner and used the others since they fitted the model better.

After the first pretest we discard all items that do not fit the model. Moreover, faulty items are corrected if the faults are minor ones, but otherwise they are also discarded. As we have constructed 125 items and we only want to save 100 for the test itself, we can throw away 25. Up till now we have not encountered any problems in getting 100 good items.

In the last construction stage the items have to be ordered according to difficulty. To achieve this we put them through a second pretest, the sample characteristics and design of which are the same as those of the first one. With the help of a Rasch analysis we calculate the item difficulties and order the items according to them. The calibrated and ordered items constitute the measuring instrument. It is at this moment not yet decided in what form precisely the items will be offered to the teachers. Probably it will be best to order them in branched tests. In that case students will only have to take a small subset of the items to have their ability level estimated. It seems to be the most economical way to use this kind of test.

In a couple of months I hope to be able to issue some more information.

Andrich, D. and Godfrey, J.R. (1979) Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: an empirical investigation using a latent trait model, *Reading Research Quarterly*, 14, 182-200.

Chapman, C.A. (1979) *A test of a hierarchical theory of reading*. The University of Chicago Press. Chicago.

Davis, F.B. (1944) Fundamental factors of comprehension in reading, *Psychometrika* 9, 185-197

Davis, F.B. (1968) Research in comprehension in reading, *Reading Research Quarterly* 3, 499-545

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Kopenhagen.

Spearrit, D. (1972) Identification of subskills of reading comprehension by maximum likelihood factor analysis, *Reading Research Quarterly* 8, 92-111

Thorndike, R. (1974) Reading as reasoning, *Reading Research Quarterly* 9, 135-147

Thurstone, L.L. (1946) Note on a reanalysis of Davis' Reading Tests, *Psychometrika* 11, 185-188